Combining Formal Concept Analysis and Translation to Assign Frames and Thematic Role Sets to French Verbs

Ingrid Falk[124] Claire Gardent[34]

[1]INRIA Nancy Grand Est

[2]Lorraine University, Nancy, France

[3]CNRS

[4]LORIA

Concept Lattices and their Applications, 2011

Summary

What this talk is about I

Use Formal Concept Analysis to associate verbs with

- syntactic information (subcategorisation frames)
- semantic information (thematic role sets)

What this talk is about II

Starting from lexical resources for French ($\sim$ a dictionary)

1. we classify French verbs based on syntactic features using FCA,
   - we filter the obtained lattice using the concept selection indices introduced in [Klimushkin et al., 2010].
   - we explore the performance of these indices, stability, separation and probability in the context of our application.
2. we extend FCA concepts with thematic role set(s) by translating English verb classes.

Summary

Verbs in Natural Language Processing Applications.

- NLP applications analyse texts to answer the question Who did What to Whom.
- They need to detect events and participants in those events.
- Events are mostly lexicalised using verbs.
- Knowledge about their syntax/semantics is crucial for NLP applications.

Syntactic and semantic information about verbs

Syntax (syntactic arguments):

John	*throws*	*the ball*	*to Mary.*
SUBJ	V	OBJ	POBJ
Agent	V	Theme	Destination

Semantics (thematic roles):

John	*throws*	*Mary*	*the ball.*
SUBJ	V	OBJ	OBJ
Agent	V	Destination	Theme

Verb Classifications...

group together verbs with similar syntactic and/or semantic behaviour.

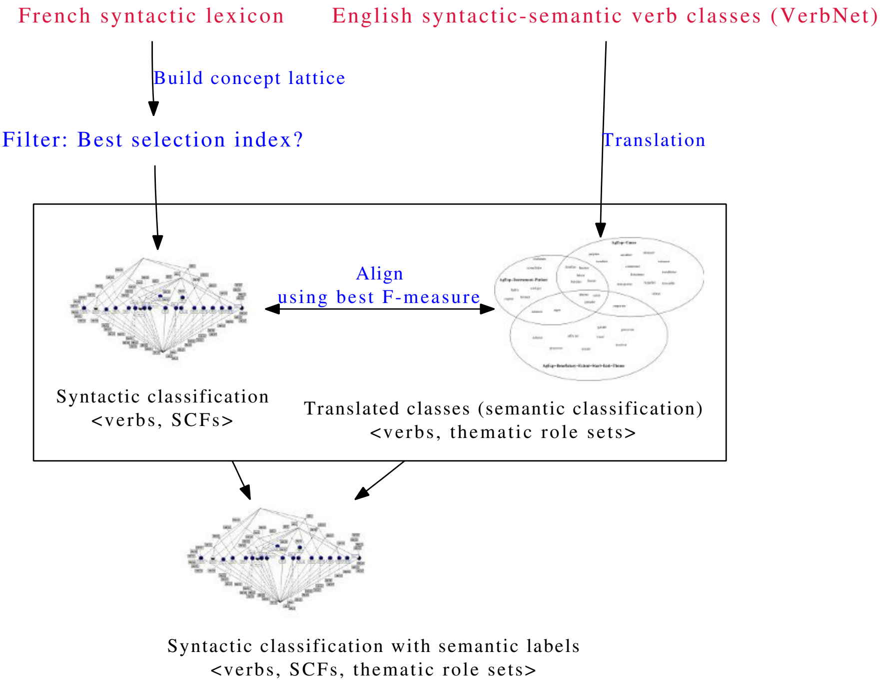
VerbNet [Schuler, 2006], example class *hit-18.1*:

Verbs:	*batter, beat, bump, butt, drum, hammer, hit, jab, kick, knock, lash, pound, rap, slap, smack, smash, strike, tap*	
Thematic roles	Agent, Instrument, Patient	
Frames	SUJ:NP,P-OBJ:PP	Agent V Patient
	SUJ:NP,P-OBJ:PP,P-OBJ:PP	Agent V Patient Instrument
	SUJ:NP,OBJ:NP	Agent V Patient
		Instrument V Patient
	SUJ:NP,OBJ:NP,P-OBJ:PP	Agent V Patient Instrument

Here we identify each VN class with its set of roles:

hit-18.1 $\rightsquigarrow$ Agent-Instrument-Patient

Summary


French syntactic lexicon
English syntactic-semantic verb classes (VerbNet)
Build concept lattice
Filter: Best selection index?
Translation
Align
using best F-measure
Syntactic classification
<verbs, SCFs>
Translated classes (semantic classification)
<verbs, thematic role sets>
Syntactic classification with semantic labels
<verbs, SCFs, thematic role sets>

Summary

Building and Filtering Verb Classes using FCA

- build a context from a syntactic lexicon of French verbs.
- compute the lattice – Galicia[1],
- filter using concept stability, separation and/or probability – [Klimushkin et al., 2010],

[1] http://www.iro.umontreal.ca/ galicia/

French subcategorisation lexicon.

Example entries

verb	frame and example
manifester	SUJ:NP, OBJ:NP Cette expression manifeste un dédain réel. This expression shows a real disdain.
manifester	SUJ:NP, OBJ:NP, A-OBJ Il ne manifeste jamais ses vrais sentiments (à qqn.) He never showed his true feelings (to sb.)

Merged from:

- Dicovalence, [van den Eynde and Mertens, 2003]
- the LADL tables, [Gross, 1975]
- TreeLex, [Kupść and Abeillé 2008]

⇝ 5918 verbs, 345 subcategorisation frames, 20 443 verb-frame pairs.

The concept lattice.

The context:

Objects: verbs from French syntactic lexicon,

Attributes: frames from French syntactic lexicon.

$\rightsquigarrow$ a context of 2091 objects (verbs) and 238 attributes (subcategorisation frames)[a].

[a] We only use a subset of the verbs.

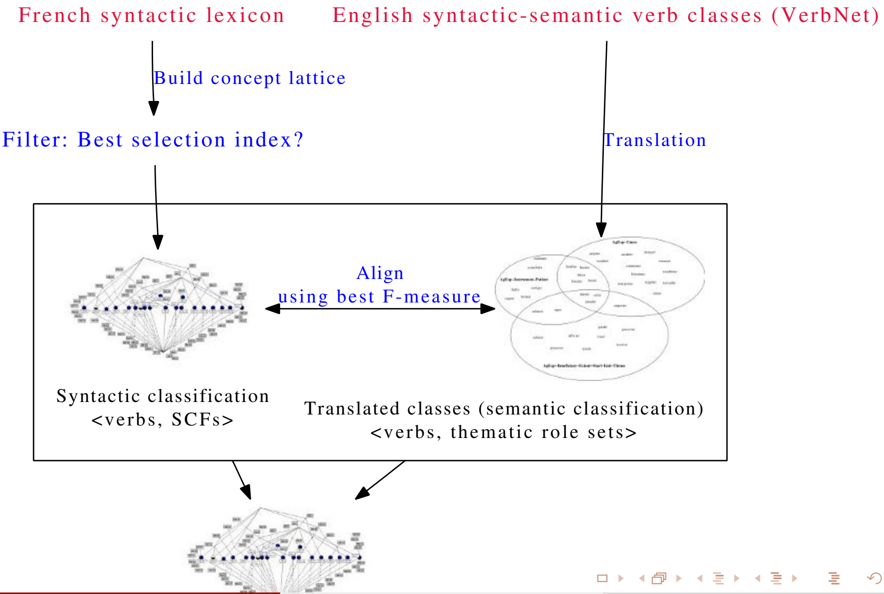
$\rightsquigarrow$

A concept lattice with 12802 concepts

Most concepts are not interesting:

- only 1 or 2 verbs,
- few frames.

How to select the most most relevant concepts?


French syntactic lexicon
English syntactic-semantic verb classes (VerbNet)
Build concept lattice
Filter: Best selection index?
Translation
Align
using best F-measure
Syntactic classification
<verbs, SCFs>
Translated classes (semantic classification)
<verbs, thematic role sets>

Concept Selection Indices

Use

- Concept stability
- Concept separation
- Concept probability

introduced in [Klimushkin et al., 2010] for selecting relevant concepts in concept lattices built on noisy data.

Concept stability

Definition ([Kuznetsov, 1990])

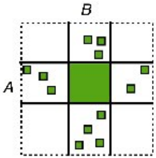
Let (V, F) be a formal concept and $'$ the derivation operator. It's intensional stability is defined as:

$$\sigma_i((V, F)) := \frac{|\{A \subseteq V \mid A' = F\}|}{2^{|V|}}$$

- the proportion of the subsets of the extent which have the same intent.
- a more stable concept is less dependant on individual members in the extension.

Concept separation

Ratio between the objects and attributes covered by the concept and the objects and attributes covered by the intent and extent of each of the concept's objects and attributes.



Definition ([Klimushkin et al., 2010])

$$\mathfrak{s}((V,F)) = \frac{|V|\,|F|}{\sum_{v \in V} |\{v\}'| + \sum_{f \in F} |\{f\}'| - |V|\,|F|}.$$

- measure about objects and attributes at same time (in contrast to stability and probability),
- concept with higher separation $\rightsquigarrow$ better sorts out verb/frames it covers from other verb/frames.

Concept Probability I

Let A be the attribute set and O the set of objects.

Probability of an arbitrary object to have attribute $a \in A$:

$$p_a = \frac{|\{a\}'|}{|O|}.$$

The probability of an arbitrary object to have all attributes from $B \subseteq A$:

$$p_B = \prod_{a \in B} p_a$$

Probability of $B \subset A$ closed:

$$p(B = B'') = \sum_{k=0}^{n} \left[\binom{n}{k} p_B^k (1 - p_B)^{n-k} \prod_{a \notin B} (1 - p_a^k) \right]$$

Concept Probability II

small $p(B = B'')$ $\rightsquigarrow$ small probability of attribute combination B to be a concept intent by chance.

$p(B = B'') \approx 1$ $\rightsquigarrow$ high probability that B is closed by chance.

But beware:

Equations based on independence of attributes – not warranted in our application!

Computing Stability, Separation and Probability

Stability
- #P complete [Kuznetsov, 2007]
- Concept lattice known ⇒ can be computed efficiently [Roth et al., 2006]
- Computations were feasible with this algorithm.

Separation
- can be computed in $\mathcal{O}(|O| + |A|)$ time, O object set, A attribute set.
- least prohibitive of three indices.

Probability
- probability of one concept: $\mathcal{O}(|O|^2 \cdot |A|)$ multiplication operations.
- we could not compute exact concept probability.
- had to use approximation!

Which works best for our application?

Which are the (combination of) indices allowing to select the best classes from our concept lattice?

Method:
Using a given (combination of) indices:

- select N concepts from concept lattice with highest index (combination),
- align these concepts with classes translated from VerbNet,
- compare obtained ⟨verb, VN class⟩ associations with a reference.

Best (combination of) indices:

- ⟨verb, VN class⟩ associations are closest to reference,
- concepts associated to VN classes cover large proportion of verbs.

Aligning concepts with translated VerbNet classes

Each translated VerbNet class C_{VN} is associated with the FCA concept C_{FCA} with best F-measure between recall R and precision P:

$$R = \frac{|\text{verbs} \in C_{\mathrm{VN}} \cap C_{\mathrm{FCA}}|}{|\text{verbs} \in C_{\mathrm{VN}}|}, P = \frac{|\text{verbs} \in C_{\mathrm{VN}} \cap C_{\mathrm{FCA}}|}{|\text{verbs} \in C_{\mathrm{FCA}}|}, F = \frac{2RP}{R+P}$$

We select the concepts with an associated translated VerbNet class C_{VN}.

These concepts group

- a set of verbs (extent of the FCA concept),
- a set of subcategorisation frames (intent of the FCA concept),
- one or more sets of thematic roles

Comparing to the reference.

Reference is set of French verbs associated to (translated from) VerbNet classes.

We compute ⟨verb, VN class⟩ pairs from both

- selected FCA concepts,
- reference.

We compute precision P, recall R and F_2-measure (because recall is more important):

$$R = \frac{|\text{pairs derived from FCA} \cap \text{pairs derived from Ref}|}{|\text{pairs derived from Ref}|}$$

$$P = \frac{|\text{pairs derived from FCA} \cap \text{pairs derived from Ref}|}{|\text{pairs derived from FCA}|}$$

$$F_2 = \frac{3RP}{R + 2P}$$

Stability, separation and probability separately I

Stability and separation: select 1500 concepts with highest index

Probability:
- low index $\rightsquigarrow$ improbable concepts (with large number of frames)
- index $\approx 1 \rightsquigarrow$ attribute combination may occur by chance
- used 6th 10 quantile ($\approx$ 1500 concepts) to assess probability separately

	cov.	**prec.**	**rec.**	F_2
stab only	39.88	18.96	32.55	26.27
sep only	34.25	28.37	21.52	23.41
prob only	35.53	26.60	20.73	22.38
w/o filtering	100	12.30	60.96	26.30

Table: F_2 scores and coverage for stability, separation and the 6th probability 10-quantile.

Stability, separation and probability separately II

- stability alone $\rightsquigarrow$ F_2 close to upper bound,
- separation and probability not suitable to be used alone,
- coverage unsatisfactory.

Results confirm observations in [Klimushkin et al., 2010].

Linear Combination

- selection index $= k_{stab} \cdot \text{stability} + k_{sep} \cdot \text{separation} - k_{prob} \cdot \text{probability}$,
- select 1500, 1000, 500 concepts with best selection index.

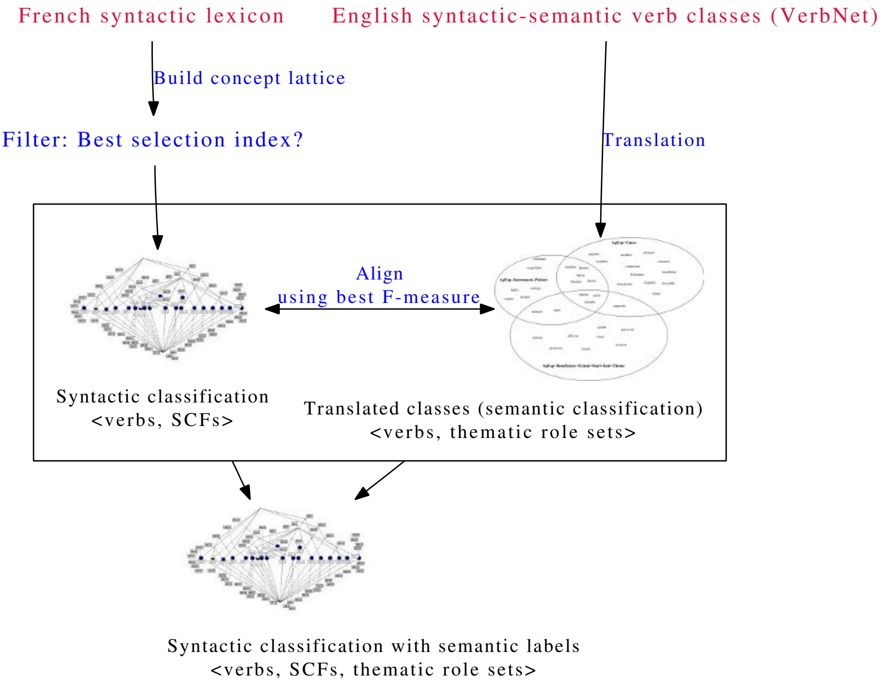
best linear combination: stability + separation, $k_{stab} = k_{sep} = 1, k_{prob} = 0$

$F_2 = 25.16$, close to upper bound, coverage 98.04%
Only $\sim$ 10% of the original lattice $\rightsquigarrow$
$\langle$verb, semantic role set$\rangle$ alignment close to alignment from entire lattice!

Other observations

- probability does not seem to have a positive effect on the selected concepts
- probability improves F_2 measure for lower number of selected concepts (1000, 500).

Summary


French syntactic lexicon
English syntactic-semantic verb classes (VerbNet)
Build concept lattice
Filter: Best selection index?
Translation
Align
using best F-measure
Syntactic classification
<verbs, SCFs>
Translated classes (semantic classification)
<verbs, thematic role sets>
Syntactic classification with semantic labels
<verbs, SCFs, thematic role sets>

Method

1) build classes grouping French verbs and SCFs using FCA
2) select 1500 concepts where *stability* + *separation* is highest
3) translate English verbs in English VerbNet classes to French (using dictionaries)
4) for each translated VerbNet class C_{VN} find the concept(s) C_{FCA} with best precision P, recall R, F-measure:
$$R = \frac{|\text{verbs} \in C_{\text{VN}} \cap C_{\text{FCA}}|}{|\text{verbs} \in C_{\text{VN}}|}, P = \frac{|\text{verbs} \in C_{\text{VN}} \cap C_{\text{FCA}}|}{|\text{verbs} \in C_{\text{FCA}}|}, F = \frac{2RP}{R+P}$$
5) associate these FCA concepts with the VerbNet class's thematic role sets and select those FCA concepts "labeled" with a thematic role set.

Effectively we obtain a classification associating:

- groups of French verbs,
- groups of subcategorisation frames,
- sets of thematic roles

What can we draw from this classification?

Concept 5312

contains verbs *bouger, déplacer, emporter, passer, promener, envoyer, expédier, jeter, porter, transmettre, transporter*

verbs can be used in construction SUJ:NP,OBJ:NP,POBJ:PP,POBJ:PP

thematic roles participating in events described by these verbs are AgExp (Agent or Experiencer), Theme, Start, End

Observations

- verbs in example are verbs of movement: an agent moves a theme from start point to end point ⇝ associations are correct,

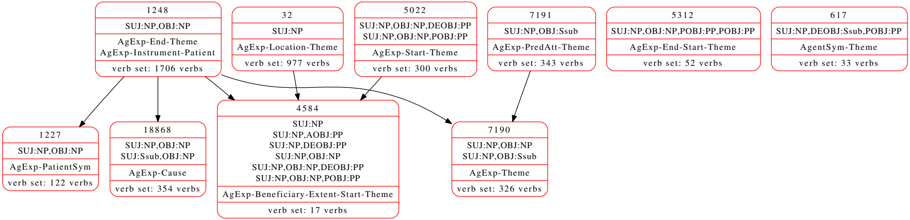


Figure: French verb $\leftrightarrow$ synt. frames $\leftrightarrow$ thematic role set associations.

Problems

- Some concepts are associated with several thematic role sets,
- Subconcepts inherit thematic role sets from superconcepts.

Many verbs belong to several VerbNet classes ...
But in what cases is the multiple mapping really warranted?

Summary

Conclusion

- We introduced a new approach of building syntactic-semantic verb classes for French based on:
  - Formal Concept Analysis on French lexical resources and
  - Translation of English syntactic-semantic resources
- We explored performance of concept selection indices stability, separation and probability [Klimushkin et al., 2010]:
  - sum of stability and separation performs best in our setting
  - selected concepts (10% of total) produced F-measure and coverage similar to when selecting from entire lattice.

Future Work

- How to use concept hierarchy relations?
- How do classifications produced by this method compare to the gold classification in the literature?
- Does the classification improve performance of a semantic role labeling (SRL) task on a corpus?
- How do classifications produced with other clustering methods perform?
  - compared to the gold classification
  - on the SRL task

Thank you!

Summary

Lexical Resources. I

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998).
The Berkeley FrameNet Project.
International Conference on Computational Linguistics, Montreal, Quebec, Canada.

Dubois, J. and Dubois-Charlier, F. (1997).
Les Verbes français.
Larousse.

Fellbaum, C., editor (1998).
WordNet: An Electronic Lexical Database.
MIT Press, Cambridge, MA.

Lexical Resources. II

Gross, M. (1975).
Méthodes en syntaxe.
Hermann, Paris.

Anna Kupść and Anne Abeillé.
Growing treelex.
In Alexander Gelbkuh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *Lecture Notes in Computer Science*, pages 28–39. Springer Berlin / Heidelberg, 2008.

Levin, B. (1993).
English Verb Classes and Alternations: a preliminary investigation.
University of Chicago Press, Chicago and London.

Lexical Resources. III

Saint-Dizier, P. (1999).
Alternation and Verb Semantic Classes for French: Analysis and Class Formation.
In *Predicative forms in natural language and in lexical knowledge bases*. Kluwer Academic Publishers.

Schuler, K. K. (2006).
VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon.
PhD thesis, University of Pennsylvania.

van den Eynde, K. and Mertens, P. (2003).
La valence: l'approche pronominale et son application au lexique verbal.
Journal of French Language Studies

Lexical Resources. IV

Sun, L., Korhonen, A., Poibeau, T., and Messiant, C. (2010).
Investigating the cross-linguistic potential of VerbNet-style classification.
In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1056–1064, Stroudsburg, PA, USA. Association for Computational Linguistics.

FCA I

Jay, N., Kohler, F., and Napoli, A. (2008).
Analysis of social communities with iceberg and stability-based concept lattices.
In *ICFCA'08: Proceedings of the 6th international conference on Formal concept analysis*, pages 258–272, Berlin, Heidelberg. Springer-Verlag.

Klimushkin, M., Obiedkov, S., and Roth, C. (2010).
Approaches to the selection of relevant concepts in the case of noisy data.
In Kwuida, L. and Sertkaya, B., editors, *Formal Concept Analysis*, volume 5986 of *Lecture Notes in Computer Science*, chapter 18, pages 255–266. Springer Berlin / Heidelberg, Berlin, Heidelberg.

FCA II

Kuznetsov, S. O. (1990).
Stability as an estimate of the degree of substantiation of hypotheses derived on the basis of operational similarity.
In *Nauchn. Tekh. Inf., Ser.2 (Automat. Document. Math. Linguist.)*, 12:21–29.

Kuznetsov, S. O. (2007).
On stability of a formal concept.
Annals of Mathematics and Artificial Intelligence, 49(1-4):101–115.

Roth, C., Obiedkov, S. A., and Kourie, D. G. (2006).
Towards concise representation for taxonomies of epistemic communities.
In *CLA*, pages 240–255.